UNLIMITED

RSRE MEMORANDUM No. 4136

ROYAL SIGNALS & RADAR ESTABLISHMENT

THE GILLICK TEST - A METHOD FOR COMPARING TWO SPEECH RECOGNISERS TESTED ON THE SAME DATA

Author: S J Cox

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
R S R E MALVERN,
WORCS.

UNLIMITED

R.S.R.E. Memorandum 4136

The Gillick Test - A Method for Comparing Two Speech Recognisers Tested on the Same Data

Stephen Cox

22nd February, 1988

Abstract

The question of the statistical significance of the difference in error rates of two speech recognition algorithms is almost invariably ignored in the literature. If it is considered, it is usually assumed that the algorithms were tested on two independent test sets, whereas in reality, they are normally tested on the same set. The Gillick Test is a simple and elegant technique for deciding whether the difference between the error rates of two algorithms tested on the same data is significant.

The author is an employee of British Telecom on secondment to RSRE. Acknowledgment is made to the Director of Research, British Telecom Research Laboratories, for permission to publish this memo.

Contents

1 Introduction

Assessment is certainly the most neglected aspect of work upon automatic speech recognition but it is vitally important. The literature currently abounds with descriptions of new algorithms and techniques for speech recognition which show an improvement over previous algorithms, but rarely, if ever, do they address the key question of whether the obtained improvement in performance is statistically significant. This memo describes a very simple test which enables comparison of two speech recognisers tested on the same set of utterances, normal practice when developing new algorithms. It is entirely based on unpublished notes by Larry Gillick of Dragon Systems Inc.

2 Some Preliminaries

The *Binomial distribution* gives the probability of exactly k errors occuring in n trials when the underlying probability of an error is e, i.e.:

$$\begin{aligned} \Pr(\text{k errors}) &= \binom{n}{k} e^k(1-e)^{n-k} \qquad k = 1, 2, \ldots, n \\ &\equiv \mathcal{B}[n, e] \end{aligned}$$

The expectation (mean) of the above Binomial distribution is ne and the variance is $ne(1-e)$. When n is large, the Binomial distribution can be approximated by a Normal distribution with mean ne and variance $ne(1-e)$, i.e.:

$$\mathcal{B}[n, e] \approx \mathcal{N}[ne, ne(1-e)]$$

A result we shall use in section 4 is: if A and B are Normally distributed random variables with expectations $E(A) = \mu_A$ and $E(B) = \mu_B$, then $E(A+B) = \mu_A + \mu_B$. Futhermore, if A and B are *independent*, $Var(A \pm B) = Var(A) + Var(B)$. Finally, it is assumed that we are dealing with recognition of isolated utterances and that no rejections are allowed (or alternatively, rejections are counted as misclassifications). Hence the error rate of the recogniser is defined to be the probability that it misclassifies an utterance.

3 Hypothesis Testing

Hypothesis testing is a standard way of quantifying the statistical significance of data produced by different processes is. A *null hypothesis* H_0, is proposed, the data is analysed and H_0 is accepted or rejected at a certain level of significance. Suppose our two speech recognisers are R_1 and R_2; the null hypothesis (H_0) is:

R_1 and R_2 have the same underlying (but unknown) error-rate.

If subsequent analysis of the data showed that we should reject H_0 at the 0.1% level, this means that if H_0 were in fact true, we would only observe a discrepancy between the error rates equal to or greater than that actually observed on 0.1% of occasions. Note that rejection of H_0 does not strictly tell us which recogniser is better, but it is safe to take the commonsense view here. A useful introduction to hypothesis testing and the use of standard tables (see next section) is given in [1].

4 Testing on Independent Test Sets

Firstly, we consider the case where R_1 and R_2 are tested on two *independent* test-sets, each of size n utterances. This introduces some of the statistical ideas which are used in section 5 when they are tested on the same data.

Suppose the underlying (but unknown) error rate of recogniser R_1 is e_1 and R_1 makes X_1 errors on its test set. Then from section 2:

$$\begin{aligned} X_1 &= \mathcal{B}[n, e_1] & (1) \\ &\approx \mathcal{N}[ne_1, ne_1(1-e_1)] & (2) \end{aligned}$$

The best estimate $\hat{e_1}$ of e_1 is:

$$\hat{e_1} = \frac{X_1}{n} \tag{3}$$

so using equations 2 and 3, $\hat{e_1}$ will be Normally distributed:

$$\hat{e_1} \approx \mathcal{N}\left[e_1, \frac{e_1(1-e_1)}{n}\right] \tag{4}$$

Similarly for R_2, which has error rate e_2:

$$\hat{e_2} \approx \mathcal{N}\left[e_2, \frac{e_2(1-e_2)}{n}\right] \tag{5}$$

The key to testing H_0 is to consider the mean and variance of the random variable $\hat{e_1} - \hat{e_2}$ [1]. Applying the result stated in section 3 for random variables A and B to equations 4 and 5 gives:

$$\hat{e_1} - \hat{e_2} \approx \mathcal{N}\left[e_1 - e_2, \frac{e_1(1-e_1)}{n} + \frac{e_2(1-e_2)}{n}\right]$$

If H_0 holds, $e_1 = e_2 = e$ (say) and:

$$\hat{e_1} - \hat{e_2} \approx \mathcal{N}\left[0, \frac{2e(1-e)}{n}\right] \qquad \text{if } H_0 \text{ holds} \tag{6}$$

The probability of observing the measured value of $\hat{e_1} - \hat{e_2}$ from a Normal distribution with zero mean and variance $2e(1-e)/n$ then tells us at what level of statistical significance to accept or reject H_0. This probability is easily found by consulting standard statistical tables. Note that in estimating the variance, e is unknown and can be estimated as $\hat{e}$ (the average estimated error) $= (X_1 + X_2)/2n$.

4.1 An Example using Independent Test Sets

Let us take an example to illustrate this. In a recent test, it was found that two recognisers R_1 and R_2 gave 72 and 62 errors respectively on a test-set of size 1400 utterances (for the purposes of this calculation, of course, we pretend that the recognisers were tested on two independent test-sets each of size 1400 utterances). The tables of the

[1] the same idea was used in [2].

cumulative Normal distribution refer to a distribution with zero mean and unity variance, so a datapoint x from a distribution with mean μ and standard deviation σ is normalised to z where:

$$z = \frac{x - \mu}{\sigma}$$

Hence we compute:

$$z = \frac{|\hat{e_1} - \hat{e_2}|}{\sqrt{\frac{2e(1-e)}{n}}} \tag{7}$$

Notice that $|\hat{e_1} - \hat{e_2}|$ is computed, because to accept or reject H_0, we are only interested in the distance of $|\hat{e_1} - \hat{e_2}|$ from zero and not whether $\hat{e_1} > \hat{e_2}$ or *vice versa*. Accordingly, we require the probability P that a point falls outside z on *either* side of the mean - this probability is shown as the shaded area in Fig 1:

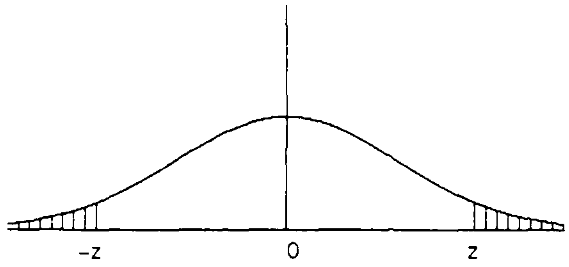

Fig 1: A two-tailed test on a Normal distribution with zero mean

We therefore use the 'two-tailed' tables of the cumulative Normal distribution, and putting the above figures into equation 7, find $z = 0.88531$ and hence $P = 0.376$ [2]. This means that if H_0 is assumed (i.e. the underlying error rates are equal), we would expect a difference between two observed error rates equal to or greater than that actually observed on 37.6 % of occasions. In other words, there is a very good chance that the underlying error rates *are* equal and all we have observed is a random effect. It will be seen in section 4.1 that the extra information provided when the recognisers are tested on the same data may greatly increase the significance of the result.

5 Testing on the Same Data Set

Consider the more realistic situation where the test set consists of a single set of utterances $U_1, U_2, \ldots, U_n$. For any utterance U_i, define the following probabilities:

$q_{00} = \Pr(R_1$ classifys U_i correctly, R_2 classifys U_i correctly)
$q_{01} = \Pr(R_1$ classifys U_i correctly, R_2 classifys U_i incorrectly)
$q_{10} = \Pr(R_1$ classifys U_i incorrectly, R_2 classifys U_i correctly)
$q_{11} = \Pr(R_1$ classifys U_i incorrectly, R_2 classifys U_i incorrectly)

[2]the NAG library function S015ABF returns $1 - \frac{P}{2}$.

These probabilities can be visualised more easily in the following table form:

		R_2 Right	R_2 Wrong
R_1	Right	q_{00}	q_{01}
	Wrong	q_{10}	q_{11}

Table 1: Joint probability of correct decision or error for two speech recognisers tested on the same data

It is clear that:

$$e_1 = q_{10} + q_{11}$$
$$e_2 = q_{01} + q_{11}$$

If H_0 holds, $e_1 = e_2$, so $q_{01} = q_{10}$. Let:

$$q = \frac{q_{10}}{q_{01} + q_{10}} \tag{8}$$

Then if H_0 holds, $q = \frac{1}{2}$. Equation 8 may be interpreted as follows: $q_{01} + q_{10}$ is the probability that only one of the recognisers makes an error on a given utterance; hence q is the probability that R_1 makes an error on a given utterance given that only one of the recognisers makes an error.

Of course the q_{xx} probabilites are computable only with an infinite test set. However, we can estimate them from our finite test set. Define:

n_{00} = No of utterances which R_1 classifys correctly, R_2 classifys correctly
n_{01} = No of utterances which R_1 classifys correctly, R_2 classifys incorrectly
n_{10} = No of utterances which R_1 classifys incorrectly, R_2 classifys correctly
n_{11} = No of utterances which R_1 classifys incorrectly, R_2 classifys incorrectly

Once again, this is more easily visualised as:

		R_2 Right	R_2 Wrong
R_1	Right	n_{00}	n_{01}
	Wrong	n_{10}	n_{11}

Table 2: Distribution of numbers of correct decisions or errors for two speech recognisers tested on the same data

Now:

$n_{01} + n_{10}$ = No of utterances on which only one recogniser makes an error
n_{10} = No of utterances on which R_1 makes an error, R_2 classifys correctly
q = Pr(R_1 makes an error given that only one recogniser makes an error)

These three statements should make it clear that:

$$\begin{aligned} n_{10} &= \mathcal{B}[n_{01} + n_{10}, q] \\ &\approx \mathcal{N}[q(n_{01} + n_{10}), q(1-q)(n_{01} + n_{10})] \end{aligned}$$

The best estimate of q is $\hat{q}$ where:

$$\begin{aligned}\hat{q} &= \frac{n_{10}}{n_{01}+n_{10}} \\ &\approx \mathcal{N}\left[q, \frac{q(1-q)}{n_{01}+n_{10}}\right]\end{aligned}$$

If H_0 holds, $q = \frac{1}{2}$, so:

$$\hat{q} \approx \mathcal{N}\left[\frac{1}{2}, \frac{1}{4(n_{01}+n_{10})}\right] \qquad \text{if } H_0 \text{ holds} \tag{9}$$

Compare equations 6 and 9. The probability of observing $\hat{q}$ from the Normal Distribution in equation 9 indicates at what level of statistical significance to accept or reject H_0. Following the steps laid out in section 4.1, compute:

$$z = \frac{\left|\hat{q} - \frac{1}{2}\right|}{\sqrt{\frac{1}{4(n_{01}+n_{10})}}} \tag{10}$$

and use the same 'two-tailed' test to determine whether to reject or accept H_0.

5.1 Examples using the Same Test Set

We can now drop the pretence of section 4.1 of two independent test sets and repeat the calculation on the basis that R_1 and R_2 were tested on the same test-set. The distribution of errors from this test was:

		R_2	
		Right	Wrong
R_1	Right	2721	7
	Wrong	17	55

Hence $n_{01} = 7$, $n_{10} = 17$. $\hat{q} = 0.70833$, $z = 2.041$ and $P = 0.0412$. If H_0 were true, error patterns indicating a discrepancy between the error rates as large as or larger than this would be observed on only just over 4% of occasions (c.f. 37.6% of occasions if independent test sets are assumed), so there is quite a good chance that a genuine difference exists.

It is instructive to compare the values of P for different error patterns. For instance, suppose that R_1 and R_2 made the same numbers of errors as above but the error pattern was:

		R_2	
		Right	Wrong
R_1	Right	2721	62
	Wrong	72	0

Here, $P = 0.3876$ so there is very little evidence for a difference between the recognisers. Consider another error pattern:

		R_2	
		Right	Wrong
R_1	Right	2721	0
	Wrong	10	62

$P = 0.00517$, convincing evidence for a difference.

5.2 Comments on the examples

Notice that n_{00} and n_{11} are not considered in the calculations, so that information on the relative performance of the classifiers is supplied only when they disagree. A large value of $|n_{10} - n_{01}| \Rightarrow$ large $\hat{q} \Rightarrow$ large z in equation 10, indicating the possibility of a genuine difference in error rates; however, z is 'tempered' by the term $1/4(n_{10} + n_{01})$ which is large when $n_{10} + n_{01}$ is small, reducing z and hence the significance of the result. These observations tie up satisfyingly with one's intuitions about testing two classifiers on the same data. It is worth mentioning that the more disjunct the error pattern is (i.e. the higher the ratio $(n_{10} + n_{01})/n_{11}$), the greater the improvement in performance obtainable by constructing a combined classifier (using some means of arbitration when R_1 and R_2 disagree).

6 Discussion and Summary

The Gillick test (actually an application of McNemar's test) puts the comparison of two classifiers tested on the same data on a firm statistical footing. A feature of the test is that it takes account only of utterances on which the classifiers disagree, an obvious (but hithereto unexploited) strategy for a comparative test. Depending on the distribution of errors, it can place a much higher statistical significance on the difference in the error rates than that given by the (almost always incorrect) assumption of two independent test sets. It is very simple to apply and it is recommended that it be used whenever two recognisers are tested on the same data.

References

[1] M.J. Moroney. *Facts from Figures.* Penguin Books Ltd., 1951.

[2] S.J. Cox. *Estimating the error rates of isolated word template matching speech recognisers.* Technical Report R18/005/86, BT Technology Executive, 1986.

DOCUMENT CONTROL SHEET

Overall security classification of sheet ..UNCLASSIFIED..

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference	3. Agency Reference	4. Report Security Classification
	Memorandum 4136		Unclassified

5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location
778400	Royal Signals and Radar Establishment St Andrews Road, Malvern, Worcestershire WR14 3PS

5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

THE GILLICK TEST - A METHOD FOR COMPARING TWO SPEECH RECOGNISERS TESTED ON THE SAME DATA

7a. Title in Foreign Language (in the case of translations)

7b. Presented at (for conference papers) Title, place and date of conference

8. Author 1 Surname, initials	9(a) Author 2	9(b) Authors 3,4...	10. Date	pp. ref.
Cox S			1988.02	7

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution statement

Unlimited

Descriptors (or keywords)

continue on separate piece of paper

Abstract

The question of the statistical significance of the difference in error rates of two speech recognition algorithms is almost invariably ignored in the literature. If it is considered, it is usually assumed that the algorithms were tested on two independent test sets, whereas in reality, they are normally tested on the same set. The Gillick Test is a simple and elegant technique for deciding whether the difference between the error rates of two algorithms tested on the same data is significant.

S80/48